

Clinical biostatistics

XLI. Hard science, soft data, and the challenges of choosing clinical variables in research

Alvan R. Feinstein, M.D.* *New Haven, Conn.*

Departments of Medicine and Epidemiology of the Yale University School of Medicine

In the exchange that occurs when clinicians and statisticians collaborate to conduct a research project, the statistician is regularly asked for a great deal of mathematical activity. If the project is being planned prospectively, one of his assignments is to determine "sample size" by estimating how many patients will be needed for the research to attain stochastic significance.¹⁷ If the project is a clinical trial, the statistician is asked to prepare a suitable arrangement and schedule for the randomized allocation of the compared treatments. After the research data have been collected (or if he is first consulted after the data are already available), the statistician is customarily given the job of performing a suitable "statistical analysis". For this job, he usually develops ways of sorting the data to display the results in various tables or graphs, and he then administers the statistical ritual used for testing the "significance" of what has been found.

During these mathematical procedures, the investigating clinician has not been idle. He is usually responsible for creating the basic ideas to be tested in the research, for assembling any additional clinicians who participated cooperatively, for recruiting the patients who volunteered themselves for observation, for providing or arranging to provide suitable care for those patients, and for recording the information that describes what happened to them. At

certain key moments when the results are evaluated, the clinician is responsible for helping decide whether a "significant" finding is substantively rather than merely stochastically significant.¹⁷

Although this division of research territory seems well adapted for the different talents, backgrounds, and interests of the clinical and statistical collaborators, a crucial scientific zone of the territory has not yet been discussed. This zone deals with the strategy used for choosing the clinical variables that will be observed during the research, and with the tactics that convert the clinical observations into analyzable data. The unsatisfactory management of these strategies and tactics is responsible for most of the inadequately designed research projects that have created so much discontent, dissension, and controversy in clinical epidemiologic investigation today.

Important clinical variables may sometimes be deliberately omitted from the research data because the clinico-statistical collaborators have decided that the data are "soft" and unworthy of attention. On other occasions, the neglect is inadvertent, because the clinician does not specifically mention the importance of the clinical variables, or because the statistician who is given the assignment of choosing variables is not made aware of what to look for. Alternatively, information about the important clinical data may be collected in raw form, but may then be mismanaged in the way the collaborators allow it to be classified, coded, and analyzed. Regardless of whether the errors arise from omission or commission, the cause of the difficulty is a *folie à deux*, in which capable

Supported by Public Health Service Grant No. HS00408 from the National Center for Health Services Research and Development.

*Professor of Medicine and Epidemiology, Yale University School of Medicine, New Haven, Conn.; Senior Biostatistician, Cooperative Studies Program Support Center, Veterans Administration Hospital, West Haven, Conn.

clinicians and capable statisticians—working diligently, protractedly, and often harmoniously—arrive at a carefully planned, meticulously executed, extensively publicized, and avidly defended clinical research project whose value is destroyed by its defects in clinical data. My purpose in this essay is to discuss the pathogenesis, prophylaxis, and therapy of these problems in choosing and analyzing clinical variables.

1. Definition of clinical variables

The first issue to be considered is the distinction between *clinical* variables and all of the other variables used to express the data that describe medical events in people. *Demographic* variables refer to such intrinsically personal features as age, gender, race, occupation, marital status, or religion. *Paraclinical* variables refer to the information obtained, via technologic procedures, in roentgenograms, in histologic or cytologic examinations, and in laboratory results for chemical, microbiologic, and electrographic data. *Therapeutic* variables contain descriptions of the dosage, duration, or other characteristics of treatment.

Data for the types of variables just cited can be collected by someone with relatively little medical training and sophistication. A good secretary (or an inanimate questionnaire or computer terminal) can conduct the interview for obtaining demographic details. A capable laboratory or radiographic technician can often supply the paraclinical data. An operating room attendant or a pharmacist can describe surgical procedures or pharmaceutical regimens.

The distinguishing characteristic of *clinical* data is that they usually require sophisticated clinical knowledge to be acquired, interpreted, and appreciated. Among such data are: *symptomatic* variables, which express the existence and severity of the patient's subjective distresses, discomforts, or other symptoms; *co-morbid* variables, which denote the concomitant occurrence of additional diseases, beyond the "main" disease under study; *chronometric* variables, which cite the duration of different clinical (or sometimes paraclinical) manifestations; and *decisional* variables, which list the reasons why the patient (or physician) preferred one course of action rather than another.

2. Collection of the 'basic' data

In ordinary medical circumstances, a physician notices all these clinical attributes of a patient and may even write notes about many or all of them in the patient's medical record. For most research projects, however, the patient's original medical record is not the official document that becomes analyzed. Instead, an excerpt of the patient's data is entered into a special format that is often called a *case report form*. The choice of what is collected on this form will determine the data that become the basic information noted in the research. Furthermore, the information noted in the case report form does not constitute the basic data that become analyzed. For most modern analyses, the collected case report data are converted, via various systems of classification and coding, into the entries that are punched on Hollerith (IBM) cards or affixed onto magnetic tape for processing with a digital computer.

For all these reasons, the so-called basic analytic data that are available on the computer cards and tape(s) of a research project may be quite far removed from the basic events that transpired during the project and from the fundamental accounts of those events. Before reaching the punched cards or tapes, the data underwent at least five major transfer points at which the original events were selectively observed and either deliberately or inadvertently "edited". These transfer points are: (1) the patient's basic observations and reports of subjectively perceived phenomena; (2) the physician's observation of reports from the patient and from other sources; (3) the physician's recording and interpretation of these reports in the original medical record; (4) the transfer of data from the medical record to the corresponding sites in the case report form; and (5) the numerical coding, for computer storage, of what is in the case report form. (A sixth transfer, which is quite important but which will not be further discussed here, occurs when a key-puncher does the work that converts the coded data into the punched cards or magnetized tapes transmitted to the computer.)

During all these acts of transference, someone must decide which data are worth preserving, so that the information is coded, stored, and available for retrieval and analysis. The

person who makes those coding decisions will obviously want to collect what is important. If that person decides that certain data must be obtained for coding, the decision will percolate back from the coding form to the case report form, from the case report form to the physician, and from physician to the doctor-patient interchange. But if crucial data are omitted from the coding procedure that prepares the information for processing, the variables will be absent from the "basic data" used in the analysis. The patient may describe his story well; the doctor may keep an excellent medical record; the case report form may be thoughtfully prepared and conscientiously submitted—but if the clinical variables are not suitably coded, they become excluded from the "basic data".

The decision about which variables are important enough to be coded and stored is therefore fundamental to the success of a research project. How is this decision made?

3. Statistical decisions about 'importance'

From a purely statistical viewpoint, several methods are available for deciding whether a variable is important. One approach, based on the standard deviation and coefficient of variation for the data of a single variable, is to attribute minimal importance to data in which minimal variation has occurred. For example, if everyone in a particular study is either 62, 63, or 64 inches tall, the coefficient of variation for *height* would be quite small; and *height* would probably not serve as an important discriminator among the patients. A second approach, based on statistical calculations of correlation coefficients for pairs of variables, is to assume that if two variables are highly correlated, one of them is probably unimportant and can be eliminated. For example, if we found a correlation coefficient of 0.99 between the variables *color of shoes* and *color of shoelaces*, we might conclude that color of shoelaces (or color of shoes) could be omitted from the subsequent analysis.

A third statistical approach is to use a multivariate mathematical process,¹² such as factor analysis or cluster analysis, that explores the inter-relationship of a series of different variables. After the data have been suitably mas-

saged and rearranged according to certain arbitrary mathematical principles, the process produces the allegedly "important" combinations of variables, which are called *factors* or *clusters*.

A fourth statistical approach is to use a different kind of multivariate mathematical process¹² for determining the relationship between a set of candidate variables, whose individual importance is to be evaluated, and a single target variable, whose importance is previously accepted. (In mathematical terms, the target variable is called *dependent* and the candidate variables, *independent*.) For this approach, a procedure such as multiple linear regression, discriminant function analysis, or multivariate stratification can be employed to note the simultaneous effect of a large number of baseline candidate variables on an outcome target event such as death or myocardial infarction. The multivariate statistical procedures will yield standardized "partial regression coefficients" or other numbers that indicate each variable's relative importance. With the aid of computers, the statistical procedure is often conducted in a "stepwise" manner that permits each variable to be evaluated at each step in the operations, with a decision made at that step to either include or reject the candidate variable in the group deemed "important."

Regardless of the inherent merit or mathematical ingenuity of these four statistical approaches, they all suffer from the same fundamental flaw. None of the procedures can be applied until *after* the data have been coded and collected. The statistical strategies can be very helpful for making decisions about the importance of what was coded—but the strategies cannot be used to determine, in advance, what data to observe, what to report, and what to code.

(The failure of official reviewers of the controversial UGDP study³⁸ to recognize the difference between true basic data and coded "basic data" is probably responsible for the continued smoldering of the controversy. The reviewing agents believe that they checked the basic data, but what they checked was only the coded data stored in the computerized tapes. Applying multiple linear regression models to the computerized data, the statistical analysts

claim to have ruled out any baseline disparities among the treated groups. Unfortunately, however, at least sixteen major clinical variables were omitted¹⁵ from the coded baseline information that was stored in the computer. The subsequent mathematical analyses of the coded data may have been sublime in conception and magnificent in execution, but with so many important variables omitted from consideration, the results have no credibility.)

Since mathematical modes of analysis can be applied only to data that have been coded, statistical strategies obviously cannot be used either to detect important data that were not coded or to demonstrate whether the coding process was properly performed. The only method by which the clinico-statistical collaborators can find, prevent, and remedy these errors is by using good judgment. By reviewing previous clinical experience, evaluating published literature, and consulting knowledgeable clinicians, the collaborators can learn what information is worth collecting.

4. Clinical judgments about 'importance'

Few statisticians are comfortable about having to make major decisions on the basis of "clinical judgment".⁸ The judgment may not be expressed in clear, specific terms and may not be accompanied by any direct supporting evidence. The evidence, when offered, may be anecdotal rather than documentary, vague rather than precise, and qualitative rather than quantitative. The clinician may be unable to cite the characteristics of the group of patients from whom the judgment was derived, or the cited characteristics may clearly indicate that the patients were nonrepresentative. If the judgment deals with a decision about the importance of a variable, the clinician may not stipulate either the particular target affected by the variable or the intellectual, scientific, or other mechanism by which the importance was discerned. Thus, the clinician may say that "return of appetite" is a "good prognostic sign", without listing the particular entity—such as survival, relief of pain, or return to gainful employment—whose outcome is being predicted; without quantifying the magnitude or likelihood of "good"; and without defining exactly what is meant by "return of appetite".

Nevertheless, despite these imprecisions, the clinicians' judgments are often accurate, valid, and compelling. When the judgments are actually solicited, expressed, coded, and converted into analyzable data, the statistician may be shocked by their cogency. For example, although oncologists receiving statistical guidance have spent several decades using a cancer's cellular type as the main variable for predicting patients' outcomes, thoughtful clinicians have regularly known that the patient's functional condition was of prime importance in prognostic judgment. When clinicians were asked to employ prognostic judgment rather than paraclinical histology in forecasting the outcome of treatment for cancer, a statistician used the results to inform his colleagues that "Doctors ain't so dumb".³¹ When the judgments were allowed deliberate expression, even in so crude a variable as "performance status", a consortium of prominent statisticians²⁷ expressed astounded surprise that the variable was prognostically "more important than histologic type, disease extension, or any of the usual information!".

A clinician's inability to provide documentary evidence for a judgmental decision is, of course, irrelevant to the question of whether or not to code and analyze the variables suggested by the judgment. If the variables are coded, the results can be analyzed, documentary evidence will be obtained, and the accuracy of the judgment will be sustained or refuted in the collected data. If the variables are not examined, clinicians and statisticians will continue to be deprived of the valuable information that might be provided. Since almost all thoughtful clinicians will readily acknowledge the importance of clinical variables, the current fashion of omitting such variables from analysis cannot be attributed to a non-recognition of their value. The fashion arises not from ignorant neglect of the data's importance, but from entrenched ideology about the data's quality. The clinical variables are rejected or shunned because they are regarded as "soft" and unreliable, and therefore unworthy of unscientific attention.

5. The creed of 'hard' and 'soft' data

The words *hard* and *soft* are constantly used during discussions of statistical data, but are

seldom specifically defined. The words do not appear in *A Dictionary of Statistical Terms*²⁴ and are not listed in the index of about 60 statistical biostatistical textbooks in which I have just conducted an *ad hoc* search. Despite the absence of a definition, the idea of hard data exerts an enormous statistical and scientific appeal: it is what “good” research should contain.

At least two different concepts of data, however, are included under the label of *hard*. One of these concepts refers to the general form of architecture used in the research structure with which the data were collected. Thus, data from a randomized, controlled, experimental clinical trial are regarded as harder than data from a non-randomized, non-experimental survey. Data from a “case control” (trihoc) survey, where the epidemiologist obtained the information directly from each person, are regarded as harder than survey data from comparisons of mortality rates in different geographic localities, where all the information is indirect, with the denominators of the rates coming from census bureaus, the numerators coming from death certificates, and the epidemiologist acting as a scorekeeper.

In most discussions of data, however, and in the rest of this essay, the idea of hardness refers to the interior decoration of the research structure, not to its architecture. In this alternative concept, the hardness of the data depends on the quality and reliability of the fundamental, raw elements of information.

One attribute that helps make data hard is that the information be acquired objectively rather than subjectively. Thus, a doctor’s observation of whether the patient has tenderness in a knee is regarded as objective and is therefore harder than the patient’s observation and report of the pain experienced in the knee when he walks. Another common component of hardness is that the observed entity should be preservable, so that it can be re-observed and checked. Thus, a roentgenogram of the knee (or a specimen of blood, urine, or tissue, or an electrocardiogram) can be saved for re-examination and is therefore harder than the patient’s symptom, or the doctor’s palpatory sensation, whose occurrence is not preservable. A third desirable characteristic of hardness is measurement on a dimensional

rather than ordinal or nominal scale. A goniometric measurement of the knee is thus harder than a description of its pain, tenderness, or roentgenographic appearance. *Height, weight, serum cholesterol, and width of Q waves* are all measured dimensionally and are therefore harder than *briskness of reflexes, urgency of sickness, and severity of dyspnea*, which are expressed in ordinal scales (such as **0, 1+, 2+, . . . or none, mild, moderate, severe**); or *histologic type of cancer*, which is expressed in a nominal scale (such as **anaplastic, epidermoid**, etc.).

These attributes of objectivity, preservability, and dimensionality occur in the specimens examined in modern clinical laboratories and allow application of the quality control procedures that make the measurements reliable and hard, but none of these three attributes is necessary for hardness. The crucial attribute for hardness is reliability, which can be attained even when the observation is subjective, non-preservable, and non-dimensional. For example, such data as *death, Caucasian, anaplastic cancer, and occlusion of left main coronary artery* are usually regarded as hard, although subjectively observed and non-dimensionally cited. A palpatory sensation, such as a ballottable patella or a hard lymph node, is nonpreservable as well as subjective and non-dimensional, but the data might be accepted as hard if a quintet (or trio) of suitable clinical experts, examining independently, all agreed on what was reported.

The need for reliable data is obviously a basic essential of science and the quest for such data is obviously laudable and desirable. Unfortunately, however, the contemporary pursuit of *hard* data has gone beyond the boundaries of necessity and desirability. In the current fashions of clinico-statistical research, the idea of *hard data* has often become a creed, rather than a goal. As a result of the creed, hard data have been excessively venerated to an extent far exceeding their inherent importance or actual reliability; and soft data have been not merely de-emphasized, but deliberately excluded or eliminated from consideration. For the clinicostatistical worshippers of this creed, soft data are not just “dirty” and sinful; they are scabrous horrors, to be expunged from civilized numeracy.

Despite the clear, uncontrovertible virtues of hard data, the state of clinical science has been deteriorated rather than improved by the associated creed, because it produces some major pitfalls that are seldom recognized or suitably evaluated for their "adverse side effects".

6. The pitfalls of the 'hard-data' creed

The devout reverence for hard data is associated with at least four major deleterious consequences: the credulous acceptance of "hard" information that is unreliable; the substitution of irrelevant or distorted hard data for important soft data; the inflation of effort and costs of clinical trials for which unnecessarily huge sample sizes have been calculated; and the production of dehumanized science in patient care.

a. The frequent unreliability of hard data.

Although most clinical chemists know that laboratory data must be constantly monitored and checked for quality control, clinicians have commonly accepted the dimensional values of paraclinical data as "hard" merely because they are dimensional. Every laboratory procedure can produce a wrong result from time to time, and for some procedures (or some laboratories), the results may be wrong more often than they are right. In a cooperative clinical trial, the opportunity for errors or inconsistencies in paraclinical data is much greater than for research conducted at a single institution, because so many different laboratories are involved. Yet almost no attention may be given to determining whether or not the hard data collected in the trial are actually reliable.

When the attention finally occurs, the investigators may be stunned by the results. For example, after the famous UGDP study had been in progress for six years the coordinating center in 1966 issued the distressed announcement³⁷ that "periodic shipments of coded standards to each of the participating laboratories has revealed rather marked variations among the clinics in the quality and reproducibility of the creatinine determinations". In the same announcement, the urinary protein measurements were noted to be even more unreliable than the creatinine determinations. These problems made the biostatistical coordinators of

the trial express doubts that the results would ever "be useful in detecting differences among the treatment groups". (Either the unreliable measurements were somehow corrected retrospectively or else the original doubts were subsequently exorcized, because the UGDP statisticians eventually proceeded²⁶ to use the creatinine and urinary protein data for major analytic comparisons).

The problems of inter-institutional variation and unreliability in paraclinical data are now often prevented in a cooperative study by arranging for all pertinent specimens to be sent to a single "central laboratory" where the test procedure is carefully standardized and checked for accuracy. The central-laboratory strategy takes care of the problems in a pre-planned cooperative study, but it does not deal with the many problems of interpreting laboratory data that are "pooled" from diverse, non-"centralized" sources in a trial or survey.

Furthermore, the central-laboratory strategy is seldom used, although particularly necessary, in the many situations where the "apparatus" that provides the data is a physician rather than an inanimate machine. Statisticians and clinicians have been extraordinarily naive in unquestioningly accepting, as hard data, the reports that are generated by pathologists and radiologists. For example, when the observer variability of pathologists has been checked for diagnosing the histopathology of cell types in cancer^{2, 5, 10, 20, 22, 25, 30, 32, 33, 35}, the results usually show so many discrepancies and inconsistencies that the histopathologic distinctions may become impossible to interpret. Undaunted by this gross unreliability, clinicians continue to plan treatment according to cellular types; and statisticians continue to accept, code, and analyze the data.

Radiologists have similar problems in consistency. Whenever radiologists have been studied for interobserver and intraobserver agreement in the interpretation of films for pulmonary lesions^{21, 39, 40}, for heart size¹⁸, for coronary arteriography^{6, 7, 42}, or for other key data³⁴, the variability has been striking. Yet roentgenographic data are often calmly and credulously accepted as hard and reliable. In some clinical trials, where eligibility for the study depended on the

interpretation of roentgenographic findings, patients who were randomized have later had to be de-admitted from the study when the roentgenographic evidence was reviewed.

Epidemiologists are even more credulous than clinicians in accepting and analyzing grossly unreliable data. Clinicians who are familiar with the sources of information at their own institution may have good grounds for being confident about quality of data produced by the radiologist, pathologist, or clinical laboratory. No scientific investigator, however, has any grounds for confidence about the diagnoses recorded as the "cause of death" in a collection of death certificates.⁹ Nevertheless, with complacent imperturbability, epidemiologists carry out extensive analyses for data based on the causes of death cited in compendia of death certificates accumulated in different countries and eras. The only "scientific standard" that epidemiologists use for "quality" in death certificate data is a quantitative requirement. If a non-specific cause of death (such as *old age*) is recorded in more than 15% of the death certificates, the data are regarded as unreliable¹.

Finally, even when the raw data are correct and presumably reliable, they are sometimes not properly employed. For example, in the UGDP trial³⁸, 69 patients were enrolled in the study despite failure to fulfill a simple quantitative criterion of glucose intolerance. Even after the original error was discovered, the patients were maintained in the study, receiving such unnecessary treatments as an invariant daily injection of insulin at fixed dosage.

Just as all that glistens is not gold, all data that are "hard" are not necessarily reliable.

b. The replacement of meaningful soft data by irrelevant or distorted hard data. As a result of hard-data ideology, the "substitution game" described by Yerushalmy⁴¹ has become more than just a popular sport in large-scale clinical trials. It has become almost a normative standard in design.

The game begins when the clinical investigator suggests getting information about an important, soft-data variable. "No", says the consultant, "we must not use anything that is so soft. Find something that is harder". The col-

laborators then agree upon a substitute that is hard, but irrelevant or distorting of what the investigator wanted to know. For example, in almost all the trials¹⁹ in which anticoagulants were used to prevent thromboembolism in patients with myocardial infarction, the statistical data never contained an assessment of whether or not the patients developed thromboembolism; the only end point under study was death. In all of the trials in which chemotherapy has been used to provide "palliation" for patients with cancer, the occurrence of palliation is almost never statistically reported for such "soft" phenomena as pain, anorexia, ability to work, and other aspects of quality of life. The "palliative" accomplishments of oncologic treatment are usually assessed only with such hard data as survival time and size of tumor.

In the celebrated UGDP investigation of vascular complications in diabetics, vascular complications were never described in a manner that would be recognizable to most clinicians¹⁵. The main end point variable was death; and "vascular complications" were defined statistically, rather than clinically, according to a medically bizarre set of calculations and quantifications. Diabetic neuropathy in the legs, for example, was measured according to the dimensional result of a biothesiometric test of vibration in the flesh pulp of an index finger. Congestive heart failure was identified only according to the patient's report of *use of digitalis*. Antecedent myocardial infarction, as a baseline entity before treatment, was never identified. Instead, patients were cited as having dimensionally specified electrocardiographic abnormalities.

Because the hard-data creed has dominated the planning of massive, federally funded clinical trials, the trials have often been distracted from their original objectives, and have often failed to clarify the main clinical issues for which the research was allegedly intended. In addition to the abundant examples provided by the celebrated UGDP study, another set of illustrations comes from the Coronary Drug Project⁴, which is the largest and costliest clinical trial that has ever hitherto been reported. The main clinical purpose of the trial was to determine whether patients with myocardial infarction are benefited by having their serum lipids

lowered. The only “benefit” assessed in the trial, however, was prevention of death, which turned out to have essentially similar rates in most of the treated and untreated patients. The investigators have never reported whether patients were benefited in terms of relief of symptoms, ability to work, and other aspects of quality of life. Furthermore, the death rates for the CDP groups have been presented only in terms of the treatment regimen to which the patients were assigned. The death rates have never been reported according to groups of patients in whom serum lipids were or were not actually lowered.

As John Tukey³⁶ has pointed out: “Far better an approximate answer to the right question . . . than an exact answer to the wrong question”.

c. The inflation of sample sizes. The demand for a hard-data end point alters many therapeutic trials from a remedial to a prophylactic goal. Thus, if the end point is relief of symptoms, the trial is remedial; if the end point is the development of a vascular complication or death, the trial is prophylactic, since success is measured according to how well the target event has been prevented. Because the occurrence rate is usually much lower for the appearance of a non-existing entity than for the change of something that already exists, the sample sizes needed for stochastic significance in a trial become greatly inflated if the end point is converted from a remedial soft-data event to a prophylactic hard-data event; or if a soft-data prophylactic event, with a relatively high rate of occurrence, is altered to an infrequently occurring hard-data event.

For example, suppose sample size is being calculated for a clinical trial intended to test a “control” vs. an “active” therapy for the treatment of stable angina pectoris. Suppose further that we will regard “success” as an outcome in which the relative results for one treatment are 30% better than for the other; that we want the α level of “significance” to be .05 and the β level to be .10; and that we have the choice of a remedial soft-data end point, such as improvement of symptoms, or a prophylactic hard-data end point, such as death.

For the soft-data end point, let us assume that 70% of the control group will improve. The rate

of improvement required for a relative increase of 30% in the treated group will then be 91%. We are now ready to calculate the sample size of each group, according to the classical formula¹³:

$$n = \left(\frac{1}{\Delta^2} \right) \{ z_{\alpha} [2\bar{p}(1-\bar{p})]^{\frac{1}{2}} + z_{\beta} [p_2(1-p_2) + p_1(1-p_1)]^{\frac{1}{2}} \}^2.$$

Using 2-sided levels for α and β we find (from conventional statistical tables) that $z_{\alpha} = 1.96$ for $\alpha = .05$, and $z_{\beta} = 1.64$ for $\beta = .10$. According to our assumptions, $p_1 = .70$, $p_2 = .91$, and $\Delta = p_2 - p_1 = .21$. Since $\bar{p} = (p_1 + p_2)/2$, $\bar{p} = .805$. We now insert these values into the formula, grind out the calculations, and find that $n = 89$ for each treated group. The total sample size required for the two groups would be 178 patients.

For the hard-data end point, let us assume that the expected fatality rate in the control group is 10%. For a 30% reduction, the rate of death in the actively treated group should be 7%. With these assumptions, $p_1 = .10$, $p_2 = .07$, $\Delta = -.03$, and $p = .085$. Using the same values of $z_{\alpha} = 1.96$ and $z_{\beta} = 1.64$, we enter the formula for n , do the calculations, and find that $n = 2237$. The total sample size required would be 4474. Thus, by using a hard-data end point, we have made the sample size 25 times larger ($= 4474 \div 178$) than would have been needed with the soft-data end point.

A believer in the clinico-statistical creed would now respond by saying, “Yes, the hard-data end point creates a massive increase in the number of patients, costs, and effort needed for the trial. But aren’t we better off spending the extra money and work to get an answer that is reliable, rather than an answer that will be equivocal or controversial because it depends on soft data?” The answer to this question is to point out, first of all, that the question—like many issues stated in ideological terms—is oversimplified. It does not take into account the basic purpose of the trial. Was the trial intended to measure improvement of patient’s status or prolongation of life? If prolongation of life was the main purpose, then the hard-data end point is obviously necessary. If the main purpose of treatment, however, was to improve the status of patients, then the use of a hard-data end point has displaced the goal of the

trial, as well as inflated its costs, efforts, and sample size.

Furthermore, the creedal question is based on the assumption that the hard-data end point is the only possible reliable end point. With that assumption, all the extra work and costs are expended for getting the enormous numbers of patients who must be entered and followed in the trial. There is an alternative approach to reliability, however. Before the trial begins, some extra work and costs can be directed at improving the “hardness” and reliability of the desired soft-data end point. The investment of effort in making a soft-data end point reliable—by methods outlined later in this essay—will be abundantly compensated by smaller sample sizes, by more clinically pertinent results, and by the opportunity to use the improved data in future investigations. If soft data are constantly rejected merely because they are soft, with no efforts made to improve their quality, the scientifically “vicious cycle” is perpetuated. One hard-data trial with bombastically inflated sample sizes merely leads to another bombastic hard-data trial. During the ever-increasing spiral of efforts and costs, the opportunity to improve the quality of clinical variables is ignored or lost.

d. The dehumanization of clinical science. Since all of the uniquely human phenomena of patients are expressed in soft clinical data, the exclusion of such data creates a biostatistical clinical science that is deliberately dehumanized.¹¹ The results that emerge from the trials do not contain information about the things that a practicing doctor and a patient might want to know in choosing treatment. The baseline clinical condition of the treated patients—in severity, co-morbidity, and chronometry—is not described in the statistical results; and the patients’ post-therapeutic clinical conditions—in comfort, function, and quality of life—are also omitted. In reviewing the published results, a doctor and patient can find out what happened to death rates per treatment, but not what clinical kinds of patients were treated, how each kind was affected by each treatment, and what was the total clinical impact on the patient or on the patient’s family.

The clinical science that emerges is hard and reliable, but is unsatisfactory as a guide to clinical

practice because of its dehumanization. The investigators have systematically excluded the distinctly human clinical features that distinguish the practice of medicine from an abstract exercise in statistics.

7. Choosing the important clinical variables

Once a decision is made to pay attention to clinical data, many variables become evident as having major importance. Some of the variables cited earlier will be recapitulated here, and a few additional ones will be added.

a. Prognostically important variables. The first group of clinical variables to be cited are those whose value has already been demonstrated in prognostication.

(1). *Types of symptoms.* Within the spectrum of a single disease, some patients will have primary symptoms, some will have secondary symptoms (or complications), and some will be asymptomatic. The patient’s baseline location in this spectrum of symptom types has been shown to have major prognostic importance⁸.

(2). *Severity of symptoms.* Prognosis also depends on the severity of the individual symptoms. For example, a patient with symptoms of intractable congestive heart failure is worse off prognostically than someone whose cardiac decompensation creates only mild, easily controlled edema. Someone who has lost only 5% of basic body weight is better off than someone who has become cachectic.

(3). *Severity of co-morbidity.* The diseases that co-exist in addition to the main disease have already been shown to have dramatic effects on the outcome of such ailments as cancer^{14, 16} and diabetes mellitus²³.

(4). *‘Performance status’.* A patient’s functional capacity to work or to carry out acts of daily living has been shown to be prognostically important, but its importance arises mainly because “performance status” depends on the severity of symptoms and severity of co-morbidity. If both of the latter variables have been coded, performance status may be a redundant variable. Furthermore, as noted later, performance status may be assessed inaccurately because it is also affected by “psychologic status”.

(5). *Chronometry.* The duration of symptoms

and paraclinical manifestations of a disease are important for demonstrating how long the disease has been present and for estimating the disease's auxometry, or rate of progression. Since slow-growing tumors produce symptoms slowly, the prognostic importance of chronometry has now been demonstrated in several cancers.^{3, 8, 14}

b. Therapeutically important variables. A second group of important variables are those that involve an assessment of the way a therapeutic intervention has been performed, and those that deal with the outcome of treatment (or some other intervention).

(1). *Performance.* This variable refers not to such misnomers as *performance status* for a patient's functional capacity, but to the skill with which an operative procedure is carried out or to the compliance²⁸ with which a patient maintains an oral pharmaceutical regimen. The results of technologically difficult surgical operations cannot be properly evaluated without suitable appraisal of the skill of the surgical team and of the quality of the associated anesthesia and recovery-room procedures. The accomplishments of oral pharmaceutical therapy require an assessment of how well the patient has complied with the prescribed instructions. Both of these important variables are regularly omitted from the data of clinical trials.

(2). *Regulation.* When a treatment is carried out for the purpose of regulating entity A in order to prevent event B, the success of the regulation for A is often ignored when the occurrence of event B is reported. For example, hypertension, hyperglycemia, and hyperlipidemia are all regulated for the purpose of preventing subsequent cardiovascular complications. Nevertheless, the occurrence of the complications is seldom reported in relation to the degree of regulation (as **excellent, good, poor**, etc.) for patients receiving hypotensive, hypoglycemic, or hypolipidemic agents in large-scale clinical trials. The complications are regularly reported for groups of patients who were assigned a treatment or who complied with the assignment, but not according to the regulation that was achieved.

(3). *Co-intervention.* This variable refers to additional treatments or other unplanned inter-

ventions that a patient received in addition to the scheduled main treatment. The co-interventions can substantially alter the effects of the main treatment and may also, without affecting the main treatment, provide clues to the existence of major co-morbidity that may otherwise be overlooked. Nevertheless, data about co-interventions may be omitted, or assembled in an incomplete manner, or tabulated in long "laundry lists"²⁶ without any clinically effective classifications and analyses.

(4). *Detection.* This variable refers to the intensity of the search and to the diagnostic criteria applied for the outcome events that follow treatment. Unless these processes of detection have been carried out equally for the patients in the compared treatment groups, the rates of the detected outcomes cannot be fairly compared. Although "double-blind" techniques are generally used to eliminate or reduce such bias in clinical observations, investigators regularly ignore the many other ways in which detection bias can occur. An example of this problem is demonstrated¹⁵ by the unequal rates at which necropsy was performed for patients who died in the UGDP study and by the "unblinded" way in which the excerpted data about dead patients were prepared for a subsequent "blind" review. Another manifestation of the problem may be the alleged increase in cholelithiasis that has been associated with lipid-lowering regimens. Since many gallstones are asymptomatic, any event that leads to an increase in cholecystography will lead to an increased detection of silent stones that would otherwise be unnoticed. If the lipid-lowering regimens also produce gastrointestinal symptoms, unrelated to cholelithiasis, the cholecystography that is ordered during the "workup" of the symptoms may reveal many such silent stones, which will then fallaciously be attributed to the treatment rather than to detection bias. Nevertheless, the intensity of the diagnostic search for gallstones is never cited when the occurrence rates of gallstones are compared.

c. Decisional variables. A different but generally neglected kind of important clinical information can be called *decisional data*: the reasons why the patient or the physician chose

one particular course of action rather than another. By inquiring about the reasons why certain decisions were made, the attending physician or the investigator can get clues to important clinical variables, or can learn additional useful information that might otherwise be omitted.

For example, suppose non-surgical treatments are being compared in patients classified as *inoperable* for a particular disease, such as cancer or coronary artery disease. The spectrum of inoperable patients contains at least three distinctly different groups: (1) those whose anatomic lesion is beyond the boundaries of surgical operability; (2) those who have an operable anatomic lesion, but who are denied surgery because of severe co-morbidity; and (3) those who have an operable lesion and who are also in otherwise excellent health, but who refuse surgery when it is offered. Since these three groups of inoperable patients have distinctively different prognoses, the usual demands of science would call for the groups to receive separate analyses of post-therapeutic response. In the clinico-statistical creed of the past few decades, however, neither the demands of science nor the logic of clinical medicine receives suitable attention. The reasons for the decision about operability are ignored; the inoperable patients are all lumped together as though they were homogeneous; and the investigators hope that somehow the process of randomization will rectify all the flaws. Since randomization cannot possibly provide clarity for information that was ignored, the flaws become embellished or disguised rather than removed.

Another type of decisional data has been discussed elsewhere⁸ under the title of the *iatrotropic stimulus*: the reason that the patient chose to seek medical attention at the time he did and from the doctor he selected. The information uncovered in response to this inquiry can often be helpful in establishing whether or not a patient is truly asymptomatic, in evaluating the severity of symptoms, and in determining which, if any, of several co-morbid diseases was responsible for the patient's main complaints.

In studies of the efficacy of diagnostic tests, the *diagnostropic stimulus*, which is the reason

why the doctor ordered a particular test, can provide valuable data. Such information can be especially helpful in determining the discriminating capacities²⁹ of the test in screening, case-finding, or differential diagnosis; in noting whether the test's results are being used for non-diagnostic purposes in reassurance or in prognostic and therapeutic decisions; and in assessing problems of detection bias.

Of the many other kinds of decisional variables, only one more will be noted here: the reason for a patient's functional limitations. The lack of suitable attention to this type of information is a defect not in statistical creeds, but in clinical history-taking. Many clinicians believe that a patient's symptoms cannot be accepted as reliable, because they seem to be too greatly affected by complex psychologic or social phenomena. For many purposes, however, the psychologic and social phenomena are quite simple, and the source of the unreliability is often the history taker, not the history giver.

Consider a patient with chronic, stable angina pectoris who says he has not worked for eight months. A naive history taker may record that the angina keeps the patient from working. A knowledgeable history taker will ask why the patient has not been working. The question has four basic types of answer: (1) the patient has engaged in occupational activities, but has stopped them because they provoke episodes of angina; (2) eight months ago the patient's doctor told him not to work, but neither he nor his doctor knows whether occupational activities would now provoke angina; (3) the patient does not get angina at work, but has stopped working because he fears it may provoke angina or worsen his general cardiac status; and (4) the patient has never had angina at work but is unemployed because of a general reduction of the labor force at his factory, and he has been unable to find a new job. Of these possible reasons for not working, only the first is truly attributable to the pathophysiologic impact of the angina. The second and third reasons are prophylactic, and the fourth is incidental.

Now consider a patient whose employment is of prophylactic origin, and who receives a surgical bypass graft for treatment of the angina pectoris. After the operation, the patient suc-

cessfully returns to work, having been enthusiastically urged to do so by the surgeon. Had the patient received such enthusiastic urging *before* the operation, he also might have successfully returned to work. The recorded data may indicate, however, only that he was not able to work before surgery and was able to do so afterward. An "improvement" inspired by enthusiastic iatromedical therapy thus becomes attributed to a surgical bypass graft, because the effects of therapy were not assessed, as they should have been, only in patients whose physical limitations are known to be pathophysiologic.

This problem in *post hoc* reasoning with inadequate data cannot be removed by randomized trials comparing medical vs. surgical therapy for angina pectoris. The problem will persist as long as the data are inadequate, no matter how the treatments are assigned or "controlled". The solution to the problem requires no intensive awareness of human psychology and no in-depth interviews about childhood, parenthood, siblinghood, or any of the other -hoods that occupy classical psychoanalysis. All that is necessary is a simple, straightforward question, asking about the reason why a particular limitation has occurred. The answer is usually also straightforward; its interpretation requires no psychiatric training; and the physical limitation can readily be classified as pathophysiologic, prophylactic, or co-incidental. The great barrier to acquiring and analyzing this information is not the vicissitudes of the human psyche. The problem is that doctors have often not been taught either how to take an effective clinical history or how to make effective use of the data.

8. Acquiring the necessary data

Even if all the foregoing demonstrations and arguments are accepted, a devotee of the clinico-statistical creed still has a major fall-back position. This last refuge is the argument that the requisite clinical data cannot be acquired, because clinicians will either be uncooperative, or, if cooperative, will be too inconsistent, imprecise, or unstandardized in the way they make their observations and record their results. The argument emerges from a pre-

judice that has many of the characteristics of a self-fulfilling prophecy. If the clinico-statistical investigators begin with the assumption that clinicians will do a poor job of observation, the forms for acquiring data are usually designed in a way that guarantees a poor job. No space may be provided for the necessary data to be recorded; or the available slots to be checked may be designed inadequately. (The case-report forms of the celebrated UGDP study, as noted in an earlier essay¹⁵ in this series, provide excellent examples of this type of problem. Even if the UGDP clinicians were prepared to provide superb reports of clinical observation, they could not do so. The data-acquisition forms would not permit it.)

The challenges of improving the quality of clinical data are extensive enough to warrant a separate dissertation, which will be reserved for a later installment. The challenges will require attention to methods of improving (rather than merely quantifying) observer variability; strategies of establishing effective indexes, scales, and criteria for classifying clinical phenomena; tactics that are applicable for "measuring" such entities as chronometry and severity; and an appreciation of the importance of evaluating transitions as transitions, rather than as two separate states of existence. All of these activities are part of a lamentably underdeveloped territory that is now in desperate need of creative clinico-statistical attention: the domain of clinimetrics.

References

1. Armstrong, B. K., Mann, J. L., Adelstein, A. M., and Eskin, F.: Commodity consumption and ischemic heart disease mortality, with special reference to dietary practices, *J. Chron. Dis.* **28**:455-469, 1975.
2. Charlson, M. E., and Feinstein, A. R.: The axometric dimension. A new method for using rate of growth in prognostic staging of breast cancer, *J. A. M. A.* **228**:180-185, 1974.
3. Coppleston, L. W., Factor, R. M., Strum, S. B., Graff, P. W., and Rappaport, H.: Observer disagreement in the classification and histology of Hodgkin's disease, *J. Natl. Cancer Inst.* **45**:731-740, 1970.
4. Coronary Drug Project. Design, methods, and baseline results, *Circulation* **47**:Suppl. 1, 1973.
5. Correa, P., O'Connor, G. T., Berard, C. W., Axtell, L. M., and Myers, M. H.: International

- comparability and reproducibility in histologic subclassification of Hodgkin's disease, *J. Natl. Cancer Inst.* **50**:1429-1435, 1973.
6. DeRouen, T. A., Murray, J. A., and Owen, W.: Variability in the analysis of coronary arteriograms, *Circulation* **55**:324-328, 1977.
7. Detre, K. M., Wright, E., Murphy, M. L., and Takaro, T.: Observer agreement in evaluating coronary angiograms, *Circulation* **52**:979-986, 1975.
8. Feinstein, A. R.: Clinical judgment, reprinted, 1974, Huntington, N. J., Robert Krieger Co.
9. Feinstein, A. R.: Clinical epidemiology. II. The identification rates of disease, *Ann. Intern. Med.* **69**:1037-1061, 1968.
10. Feinstein, A. R., Gelfman, N. A., and Yesner, R., with collaboration of Auerbach, O., Hackel, D. B., and Pratt, P. C.: Observer variability in histopathologic diagnosis of lung cancer, *Am. Rev. Respir. Dis.* **101**:671-684, 1970.
11. Feinstein, A. R.: The need for humanised science in evaluating medication, *Lancet* **2**:421-423, 1972.
12. Feinstein, A. R.: Clinical biostatistics. XXI. A primer of concepts, phrases, and procedures in the statistical analysis of multiple variables, *CLIN. PHARMACOL. THER.* **14**:462-477, 1973.
13. Feinstein, A. R.: Clinical biostatistics. XXXIV. The other side of 'statistical significance': Alpha, beta, delta, and the calculation of sample size, *CLIN. PHARMACOL. THER.* **18**:491-505, 1975.
14. Feinstein, A. R., Schimpff, C. R., and Hull, E. W., with the technical assistance of Bidwell, H. L.: A reappraisal of staging and therapy for patients with cancer of the rectum. I. Development of two systems of staging, *Arch. Intern. Med.* **135**:1441-1453, 1975.
15. Feinstein, A. R.: Clinical biostatistics: XXXV. The persistent clinical failures and fallacies of the UGDP study, *CLIN. PHARMACOL. THER.* **19**:78-93, 1976.
16. Feinstein, A. R., Schimpff, C. R., Andrews, Jr., J. F., and Wells, C. K.: Cancer of the larynx: A new staging system and a re-appraisal of prognosis and treatment, *J. Chron. Dis.* **30**:277-305, 1977.
17. Feinstein, A. R.: Clinical biostatistics. XL. Stochastic significance, consistency, apposite data, and some other remedies for the intellectual pollutants of statistical vocabulary, *CLIN. PHARMACOL. THER.* **22**:113-123, 1977.
18. Frieden, J., Shapiro, J. H., and Feinstein, A. R.: Radiologic evaluation of heart size in rheumatic heart disease. Studies in young patients, *Arch. Intern. Med.* **111**:44-50, 1963.
19. Gifford, R. H., and Feinstein, A. R.: A critique of methodology in studies of anticoagulant therapy for acute myocardial infarction, *N. Engl. J. Med.* **280**:351-357, 1969.
20. Gilles, F. H., Winston, K., Fulchiero, A., and Leviton, A.: Histologic features and observational variation in cerebellar gliomas in children, *J. Natl. Cancer Inst.* **58**:175-181, 1977.
21. Herman, P. G., Gerson, D. E., Hessel, S. J., Mayer, B. S., Watnick, M., Blesser, B., and Ozonoff, D.: Disagreements in chest roentgen interpretation, *Chest* **68**:278-282, 1975.
22. Holmquist, N. D., McMahan, C. A., and Williams, O. D.: Variability in classification of carcinoma in situ of the uterine cervix, *Arch. Pathol.* **84**:334-345, 1967.
23. Kaplan, M. H., and Feinstein, A. R.: The importance of classifying initial co-morbidity in evaluating the outcome of diabetes mellitus, *J. Chron. Dis.* **27**:387-404, 1974.
24. Kendall, M. G., and Buckland, W. R.: A dictionary of statistical terms, ed. 3, Edinburgh, 1971, Oliver & Boyd, Ltd.
25. Kopf, A. W., Mintzis, M., and Bart, R. S.: Diagnostic accuracy in malignant melanoma, *Arch. Dermatol.* **111**:1291-1292, 1975.
26. Miller, M., Knatterud, G. L., Hawkins, B. S., and Newberry, W. B., Jr.: A study of the effects of oral hypoglycemic agents on vascular complications in patients with adult-onset diabetes. VI. Supplementary report on nonfatal events in patients treated with tolbutamide, *Diabetes* **25**:1129-1153, 1976.
27. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., et al.: Design and analysis of randomized clinical trials requiring prolonged observation of each patient, II. Analysis and examples, *Br. J. Cancer* **35**:1-39, 1977.
28. Sackett, D. L., and Haynes, R. B., editors: Compliance with therapeutic regimens, Baltimore, Md., 1976, John Hopkins University Press.
29. Sackett, D. L., and Holland, W. W.: Controversy in the detection of disease, *Lancet* **2**:357-359, 1975.
30. Saksela, E., and Rintala, A.: Misdiagnosis of prepubertal malignant melanoma, *Cancer* **22**:1308-1314, 1968.
31. Schneiderman, M. A.: Doctors ain't so dumb, *Amer. Sci.* **49**:250A, 1961.
32. Siegler, E. E.: Microdiagnosis of carcinoma in situ of the uterine cervix: A comparative study of pathologists' diagnoses, *Cancer* **9**:463-469, 1956.
33. Sissons, H. A.: Agreement and disagreement between pathologists in histologic diagnosis, *Postgraduate Med. J.* **51**:685-689, 1975.
34. Smith, M. J.: Error and variation in diagnostic radiology, Springfield, Ill., 1967, Charles C. Thomas, Publisher.
35. Traux, H., Barnett, R. N., Hukill, P. B., Campbell, P. C., and Eisenberg, H.: Effect of inaccurate pathological diagnosis on survival statistics for melanoma: Survey of cases in the

- Connecticut Tumor Registry, *Cancer* **19**:1543-1547, 1966.
36. Tukey, J. W.: The future of data analysis, *Ann. Math. Statist.* **33**:1-67, 1962.
 37. UGDP: The relation of treatment of diabetes mellitus to the development of vascular disease. A six year progress report, submitted to the National Institute of Arthritis and Metabolic Diseases, July 30, 1966.
 38. University Group Diabetes Program: A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. Part I. Design, methods, and baseline characteristics,; Part II. Mortality results, *Diabetes* **19**:(Suppl. 2):747-830, 1970.
 39. Weitzman, S., Pocock, W. A., Hawkins, D. M., and Barlow, J. B.: Observer variation in radiological assessment of pulmonary vasculature, *Br. Heart J.* **36**:280-290, 1974.
 40. Yerushalmy, J.: Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques, *Pub. Health Rep.* **62**:1432-1449, 1947.
 41. Yerushalmy, J.: On inferring causality from observed associations, *in* Ingelfinger, F. J., Relman, A. S., and Finland, M., editors: *Controversy in internal medicine*, Philadelphia, 1966, W. B. Saunders Co.
 42. Zir, L. M., Miller, S. W., Dinsmore, R. E., Gilbert, J. P., and Harthorne, J. W.: Interobserver variability in coronary angiography, *Circulation* **53**:627, 1976.